

解析 DeepMind 采用双 Q 学习 (Double Q-Learning) 深度强化学习技术



朱小虎Neil (/u/696dc6c6f01c) [+ 关注](#)

2016.02.25 18:57* 字数 1214 阅读 8683 评论 0 喜欢 5

(/u/696dc6c6f01c)

Neil Zhu, 简书ID Not_GOD, University AI 创始人 & Chief Scientist, 致力于推进世界人工智能化进程。制定并实施 UAI 中长期增长战略和目标, 带领团队快速成长为人工智能领域最专业的力量。

作为行业领导者, 他和UAI一起在2014年创建了TASA (中国最早的人工智能社团), DL Center (深度学习知识中心全球价值网络), AI growth (行业智库培训) 等, 为中国的人工智能人才建设输送了大量的血液和养分。此外, 他还参与或者举办过各类国际性的人工智能峰会和活动, 产生了巨大的影响力, 书写了60万字的人工智能精品技术内容, 生产翻译了全球第一本深度学习入门书《神经网络与深度学习》, 生产的内容被大量的专业垂直公众号和媒体转载与连载。曾经受邀为国内顶尖大学制定人工智能学习规划和教授人工智能前沿课程, 均受学生和老师好评。

原文 (<https://link.jianshu.com?t=http://arxiv.org/pdf/1509.06461v3.pdf>)

背景

为了解决序列决策问题, 我们可以学习每个行动的最优值的估计, 即采取该行动并根据后续最优策略的未来回报的期望和。在一个给定策略 π , 在状态 s 的行动 a 的真实值为

$$Q_{\pi}(s, a) \equiv \mathbb{E} [R_1 + \gamma R_2 + \dots | S_0 = s, A_0 = a, \pi],$$

其中 $\gamma \in [0, 1]$ 是折扣因子。最优值就是 $Q^*(s, a) = \max_{\pi} Q_{\pi}(s, a)$ 。最优策略就可以通过在每个状态选择最高值的行动给出。

最优行动之的估计可以通过 Q-学习获得 (Watkin, 1989), 这也是一种形式的 td 学习 (Sutton, 1988)。大多数有趣的问题涉及遍历的状态空间太大使得学习所有的行动值难以进行。所以, 我们可以通过学习一个参数化的值函数 $Q(s, a; \theta_t)$ 。标准的 Q-学习在状态 S_t 下进行行动 A_t 更新参数, 观察及时回报 R_{t+1} 和结果状态 S_{t+1} 变成:

$$\theta_{t+1} = \theta_t + \alpha(Y_t^Q - Q(S_t, A_t; \theta_t)) \nabla_{\theta_t} Q(S_t, A_t; \theta_t). \quad (1)$$

其中 α 就是标量的步长, 目标 Y_t^Q 定义如下:

$$Y_t^Q \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \theta_t). \quad (2)$$



类似于 SGD，将当前的值 $Q(S_t, A_t; \theta_t)$ 更新为目标值 Y_t^Q

深度 Q 网络

深度 Q 网络 是多层神经网络，给定状态 s 输出一个行动值得向量 $Q(s, \cdot; \theta)$ ，其中 θ 是网络的参数。对于一个 n -维状态空间和一个包含 m 个行动的行动空间，该神经网络是从 R^n 到 R^m 的映射。DQN 算法的两个最重要的特点是目标网络 (target network) 和经验回顾 (experience replay)。目标网络，其参数为 θ^- ，其实除了其参数每 τ 次从在线网络复制外都和在线网络相同，所以 $\theta^-_t = \theta_t$ ，在其他步都是固定大小。DQN 使用的目标就是：

$$Y_t^{\text{DQN}} \equiv R_{t+1} + \gamma \max_a Q(S_{t+1}, a; \theta_t^-). \quad (3)$$

对经验回顾，观察到的转换被存放一段时间，并会均匀地从记忆库采样来更新网络。目标网络和经验回顾都能大幅提升算法的性能 (Mnih et al., 2015)。

双 Q-学习

公式 (2) 和 (3) 中，在标准的 Q-学习和 DQN 中的 \max 操作使用同样的值来进行选择和衡量一个行动。这实际上更可能选择过高的估计值，从而导致过于乐观的值估计。为了避免这种情况的出现，我们可以对选择和衡量进行解耦。这其实就是双 Q-学习 (van Hasselt, 2010)。

最初的双 Q-学习算法中，两个值函数通过将每个经验随机更新两个值函数中的一个，这样就出现了两个权重集合， θ 和 θ' 。对每个更新，一个权重集合用来确定贪心策略，另一个用来确定值。为了更好地比较这两者，我们可以将 Q-学习中的选择和衡量分解，将 (2) 重写为

$$Y_t^Q = R_{t+1} + \gamma Q(S_{t+1}, \arg\max_a Q(S_{t+1}, a; \theta_t); \theta_t).$$

双 Q-学习误差可以被写成：

$$Y_t^{\text{DoubleQ}} \equiv R_{t+1} + \gamma Q(S_{t+1}, \arg\max_a Q(S_{t+1}, a; \theta_t); \theta'_t). \quad (4)$$

注意到在 $\arg\max$ 中行动的选择仍旧取决于在线的权重 θ_t 。这表示，如同 Q-学习中那样，我们仍然会根据当前值来估计贪心策略的值。然而，我们使用了第二个权重集合 θ'_t 来公平地衡量这个策略的值。第二个权重的集合可以对称式地通过交换 θ 和 θ' 的更新。

您的支持可以鼓励作者写出更多的文章。

赞赏支持





朱小虎Neil (/u/696dc6c6f01c) ♂

写了 480552 字，被 3725 人关注，获得了 2068 个喜欢
(/u/696dc6c6f01c)

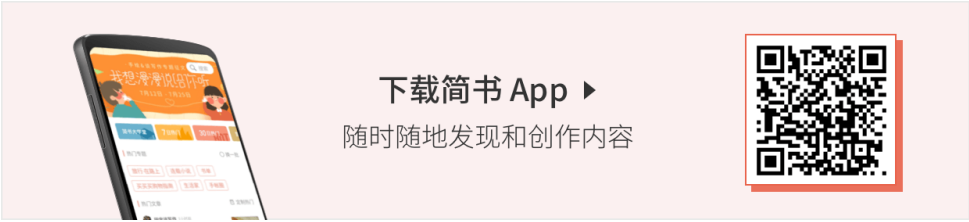
+ 关注

我是 Neil 朱小虎，University AI 创始人 & Chief Scientist，UniversityAI-AI-Unconference Meetup 组织者，...

喜欢 | 5



更多分享



(/apps/redirect?utm_source=note-bottom-click)



登录 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-comr) 发表评论

评论

智慧如你，不想发表一点想法 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-nocomments-text)咩~

被以下专题收入，发现更多相似内容



理科生的果壳 (/c/27e6af2a38aa?

utm_source=desktop&utm_medium=notes-included-collection)




DeepMind (/c/fc7215c41d48?utm_source=desktop&utm_medium=notes-included-collection)

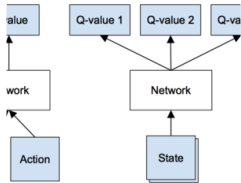
高考3500 (/p/0bda5d804ee3?utm_campaign=maleskine&utm_content=...

A a (an) [ə, eɪ(ə)n] art. 一 (个、件.....) abandon [ə' bændən] v.抛弃，舍弃，放弃 ability [ə' bɪlɪti] n. 能力；才能 able [' eɪb(ə)] a. 能够；有能力的 abnormal [æb' nɔ:m...



 o涂桃子 (/u/02eb49244585?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio


(/p/d347bb2ca53c?



(/apps/redi
utm_sourc
banner-clic

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
解析 DeepMind 深度强化学习 (Deep Reinforcement Learning... (/p/d347...

Neil Zhu, 简书ID Not_GOD, University AI 创始人 & Chief Scientist, 致力于推进世界人工智能化进程。制定并实施 UAI 中长期增长战略和目标, 带领团队快速成长为人工智能领域最专业的力量。作为行业领导者...


 朱小虎Neil (/u/696dc6c6f01c?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/aefe80044463?

- Introduction to Deep Learning
- Introduction to Reinforcement Learning
- Value-Based Deep RL
- Policy-Based Deep RL
- Model Based Deep RL


utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
深度强化学习 (/p/aefe80044463?utm_campaign=maleskine&utm_conte...

Neil Zhu, 简书ID Not_GOD, University AI 创始人 & Chief Scientist, 致力于推进世界人工智能化进程。制定并实施 UAI 中长期增长战略和目标, 带领团队快速成长为人工智能领域最专业的力量。作为行业领导者...

 朱小虎Neil (/u/696dc6c6f01c?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio


RLDM 强化学习教程-David Silver (/p/782cd2518d0e?utm_campaign=m...

Neil Zhu, 简书ID Not_GOD, University AI 创始人 & Chief Scientist, 致力于推进世界人工智能化进程。制定并实施 UAI 中长期增长战略和目标, 带领团队快速成长为人工智能领域最专业的力量。作为行业领导者...

 朱小虎Neil (/u/696dc6c6f01c?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio


无标题文章 (/p/3d905123967a?utm_campaign=maleskine&utm_content...

人工智能很可能导致人类的永生或者灭绝, 而这一切很可能在我们的有生之年发生。上面这句话不是危言耸听, 请耐心的看完本文再发表意见。这篇翻译稿翻译完一共三万五千字, 我从上星期开始翻, 熬了好几个...

 湾千 (/u/aed952c82c11?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

忘不掉的初恋 (/p/5a88c7c281dc?utm_campaign=maleskine&utm_conte...

记得上次猜数字游戏真心话大冒险, 刘金林问我谈过几个女朋友, 我说三个, 其实只有两个。一个初恋, 一个是现在的妻子。每一个个男人都不会忘记自己的初恋, 我也不例外。她是我表妹同父异母的姐姐, 我们...

 苏克雷 (/u/567955f24a31?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/ed30b4595c68?



(/apps/redi
utm_sourc

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
擅长摔跤，7段黑带柔道高手，会武术的领导人！ (/p/ed30b4595c68?utm_...

在众多国家领导人中，我最喜欢普京，一个会武术的国家领导人。普京特工出身，他从11岁就喜欢上了摔跤和拳击，但不久鼻梁骨被打断，也没有去做手术，他自认为会自己长好，果然自动痊愈了，神奇！从中学...

民间老拳师 (/u/b863cbba137f?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

致操蛋的爱情 (/p/3e6dde906e45?utm_campaign=maleskine&utm_conte...

“你还相信爱情吗？”“爱情？我们一定要在第一次正式约会的时候讨论这种话题吗？这会让我的中指勃起哎。”在被阴差阳错的相亲对象频繁骚扰后，小北终于用这句话将他拒之于千里之外。A掉了这顿饭，小北不出意...

曜黎 (/u/ddd91766e86d?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/0536a91540c8?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio
忆墨庄 (/p/0536a91540c8?utm_campaign=maleskine&utm_content=not...

春天来了，每天穿梭在潇湘大道上看草长莺飞。遇上天气好，散了学的孩童，三三两两，跑到河边上放风筝。在我心里，长沙最美的道路当属潇湘大道。清晨，河东的太阳矮矮地挂在楼宇之间，摩天大楼的玻璃...

蝙蝠君 (/u/2dd8935b853f?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

03-偏好设置 (/p/d32bcbc3a2ba?utm_campaign=maleskine&utm_conte...

AlanGe (/u/28cfb833134f?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

